Generative Data Augmentation for Improving Semantic Classification

Shadman Rohan[1*] **Mahmud Elahi Akhter**[4*] **Ibraheem Muhammad Moosa**[3]
Nabeel Mohammed[2] **Amin Ahsan Ali**[1] **AKM Mahbubur Rahman**[1†]

[1]Center for Computational & Data Sciences [2]North South University, Bangladesh
[3]Pennsylvania State University, USA [4]Queen Mary University of London, UK
shadmanrohan@gmail.com m.akhter@qmul.ac.uk ibraheem.moosa@psu.edu
nabeel.mohammed@northsouth.edu {aminali, akmmrahman}@iub.edu.bd

Abstract

We study sentence-level generative data augmentation for Bangla semantic classification across four public datasets and three pretrained model families (BanglaBERT, XLM-Indic, mBERT). We evaluate two widely used, reproducible techniquesparaphrasing (mT5-based) and round-trip backtranslation (BnEnBn)and analyze their impact under realistic class imbalance. Overall, augmentation often helps, but gains are tightly coupled to label quality: paraphrasing typically outperforms backtranslation and yields the most consistent improvements for the monolingual model, whereas multilingual encoders benefit less and can be more sensitive to noisy minority-class expansions. A key empirical observation is that the neutral class appears to be a major source of annotation noise, which degrades decision boundaries and can cap the benefits of augmentation even when positive/negative classes are clean and polarized. We provide practical guidance for Bangla sentiment pipelines: (i) use simple sentence-level augmentation to rebalance classes when labels are reliable; (ii) allocate additional curation and higher inter-annotator agreement targets to the neutral class. Our results indicate when augmentation helps and suggest that data qualitynot model choice alonecan become the limiting factor.

1 Introduction

Semantic classification is an important task in natural language processing, facilitating the understanding of textual content by machines. Among the myriad languages worldwide, Bangla stands out as one of the most spoken, yet it encounters challenges related to data scarcity and resource limitation. It is important to note that, many efforts were undertaken to create datasets for sentiment analysis in Bangla but very few meet the rigorous quality benchmarks such as measuring Inter-Annotator Agreement (IAA). In the literature of data augmentation, word replacement with synonyms or nearest embedding has been a traditional strategy. However, its efficacy is limited, especially in the context of complex languages and nuanced semantics. On the other hand, generative data augmentation provides an avenue to produce more diverse and contextually relevant data samples. Paraphrasing and backtranslation emerge as notable generative strategies, with the latter being a staple in the machine translation domain.

In this study, we show how generative data augmentation can be used to improve the performance of sentiment classification. For this purpose we use four different datasets and three different models. We show our results for both multilingual and monolingual models. We found that monolingual models along with paraphrased data augmentation performed best, followed by backtranslation. Furthermore, the multilingual language models performance did not improve with data augmentation. While improving variation of good datapoints, we found that these techniques can also magnify the noise in the datasets. Additionally, we also found considerable label noise in the datasets and we empirically showed how this noise impeded the performance of sentiment classifiers. In this paper we study the potential of generative data augmentation, with a specific focus on its application in Bangla semantic classification.

Our goal is not to propose a new augmenta-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 347–356

tion algorithm. Rather, we provide a controlled, Bangla-centric study across four public datasets and three pretrained model families (monolingual, Indic-multilingual, massively multilingual). We quantify when simple sentence-level augmentations help and, crucially, show that benefits are bounded by *label quality*, with the neutral class emerging as the primary source of noise. In summary, our contributions are primarily three-fold:

1. We investigate the viability of generative data augmentation for Bangla sentiment classification across four public Bangla social media datasets.

2. We find that paraphrasing works better as a generative data augmentation method compared to backtranslation and this effect is more prominent in monolingual language models.

3. We further identify that the neutral class is the main source of label noise in the chosen datasets.

2 Related Work

There have been numerous foundational works that have shaped the field of Sentiment Analysis (SA). One of the earliest notable researches was conducted by Hu and Liu (2004), where they presented the Aspect-Based Opinion Mining model, a predecessor of the Feature-Based Opinion Mining Model. Their work emphasized the nuances of customer reviews, providing an analytical framework for subsequent studies.

Benchmark datasets play a significant role in the evolution and validation of SA techniques. Prominent among them are the Stanford Sentiment Treebank (Socher et al., 2013), IMDB Movie Reviews Dataset (Maas et al., 2011), Amazon Product Data, and Sentiment140 (Go et al., 2009). Over the years, these datasets have served as standard benchmarking resources for sentiment analysis.

2.1 Sentiment Analysis in Bangla

When it comes to studies specific to Bangla, many prioritized the creation of clean datasets with extensive pre-processing, as seen in the works of Khatun and Rabeya (2022) and Islam et al. (2020). Furthermore, some studies, like Hossain et al. (2020), have even incorporated code-mixed samples integrating both Bangla and English. In contrast, others such as Islam et al. (2021) have made efforts to represent real-world data by incorporating noisy samples from social media platforms. In terms of classification, most researches have leaned towards a tri-class labeling system - positive, negative, and neutral. However, only a selected few, such as Islam et al. (2021), delved deeper by further dividing the positive and negative sentiments based on their intensity - weak or strong. Other techniques, such as the one presented by Abu Taher et al. (2018), involved N-Gram Based Sentiment Mining using Support Vector Machine. Chakraborty et al. (2022) explored a ternary sentiment classification for Bangla text using both Support Vector Machine and Random Forest Classifier.

Expanding the lexicon for sentiment analysis in Bangla was the primary goal for studies like (Naim, 2021) and (Bhowmik et al., 2022). An enriched vocabulary set has been observed to enhance the efficacy of sentiment models. Additionally, these studies employed aspect-based sentiment analysis techniques, a testament to the influence of Hu and Liu's foundational work. Some researches ventured into cross-lingual approaches, such as the work of Sazzed (2020), which involved translating Bangla sentences into English for sentiment analysis. However, the community did not wholly embrace this due to the dependencies it introduced. Furthermore, certain datasets were inhibited by class imbalances, as noted with Wahid et al. (2019), which had a disproportionate number of positive or negative samples compared to neutral ones. Recent works, including Islam et al. (2021) and Hossain et al. (2020), have not only emphasized the importance of data quality but also focused on the meticulousness of data annotation. Related Bangla resource creation includes BenCoref for coreference, highlighting annotation design and cross-domain coverage (Rohan et al., 2023).

Progressing further, Bhowmick and Jana (2021) applied transformer-based models like BERT and XLM-ROBERTA, signifying the continuous evolution and adaptability of sentiment analysis techniques in Bangla. Taking a more complex approach, Rafi-Ur-Rashid et al. (2022) employed an ensemble of deep learning models. This research also tackled class-imbalanced data using

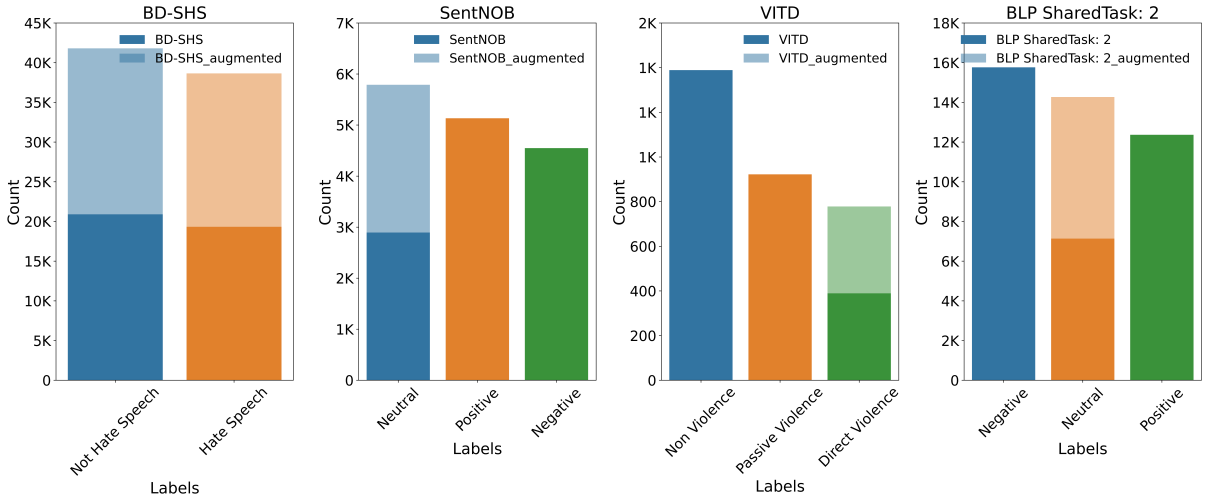

Figure 1: Comparison of dataset sizes before and after augmentation. The augmentation can be either backtranslation or paraphrasing.

the focal loss function. Despite the vast array of datasets available, many suffer from a domain bias. The SentNoB dataset (Islam et al., 2021) is currently considered the most robust, offering diversity across 13 domains and presenting a dataset with 15,000 Bangla samples.

3 Methodology & Experiment Details

We study two sentence-level augmentation methods, paraphrasing and back-translation, to increase minority-class coverage while preserving labels in Bangla sentiment analysis.

Traditional augmentation in NLP often relied on *synonym substitution* and *embedding-based replacement*. For temporal expressions, Kolomiyets et al. (2011) substituted tokens with WordNet (Miller et al., 1990) or model-predicted synonyms, assuming local replacements largely preserve meaning. Later, Kobayashi (2018) and Wei and Zou (2019) replaced words with nearest-neighbor or LM-suggested alternatives. However, in sentiment analysis even minimal lexical edits can flip polarity, and single-word substitutions frequently act like adversarial artifacts that distort semantics and inject label noise (Alzantot et al., 2018; Kaushik et al., 2020). Surveys further note that word-level noising yields limited *semantic* diversity and can harm label fidelity, recommending sentence-level methods when preservation matters (Feng et al., 2021). Accordingly, we exclude synonym/embedding replacement in our study.

3.1 Paraphrasing

Paraphrase-based augmentation aims to increase lexical and syntactic diversity while preserving task-relevant semantics, mitigating overfitting to surface forms and reducing minority-class sparsity (Fadaee et al., 2017; Wei and Zou, 2019). We generate one paraphrase per source sentence using the mT5-based Bangla paraphrasing model from Akil et al. (2022), built on Xue et al. (2021). We do not apply post-generation filtering; given the reported semantic fidelity of this resource, capping at a single paraphrase limits distributional shift (Akil et al., 2022).

3.2 Back-translation

Back-translation rewrites a sentence by translating it to a pivot language and back, typically preserving meaning while introducing natural syntactic variation (Sennrich et al., 2016; Edunov et al., 2018; Xie et al., 2020). We employ Bangla→English→Bangla using the Bangla–English MT system of Hasan et al. (2020), choosing English as the pivot due to model maturity and availability. Although round-trip translation can introduce minor noise, the resulting variation is generally label-consistent and improves robustness to social-media artifacts (informality, misspellings, code-mixing) in our datasets (Edunov et al., 2018; Xie et al., 2020).

To keep the pipeline simple and reproducible-and to isolate the effect of class rebalancingwe intentionally avoid perplexity- or classifier-based filtering. Filtering/selection strategies (e.g.,

Original	Backtranslated	Paraphrased
ইরানের একটাই শক্তি , নিজস্ব প্রযুক্তির উন্নত সব মিছাইল ব্যবস্থা	ইরানের একটা শক্তি আছে, তার নিজস্ব প্রযুক্তি উন্নত মিসিল।	ইরানের একমাত্র শক্তি, তার নিজস্ব প্রযুক্তির সব উন্নত মিচাইল সিস্টেম।
এই মেয়েদেরকে কোটি করে টাকা দিয়ে দিন , তারা তাদের পরিবার নিয়ে সুখে থাকুক	এই মেয়েদের লক্ষ লক্ষ টাকা দিন, তারা তাদের পরিবারের সাথে সুখী হবে।	এই মেয়েদের কোটি কোটি টাকা দিন, তাদের পরিবারের সঙ্গে সুখী থাকতে দিন।
পি বি আই প্রধানকে দেখে মনে হোল উনি নাটক করছেন , হায়রে দেশ	পিবিআই প্রধানকে দেখে মনে হচ্ছে সে খেলছে, আমার দেশ।	পিবিআই প্রধানকে দেখে মনে হচ্ছে সে নাটক করছে, হায় দেশ।

Figure 2: Comparison of semantic preservation between paraphrased and backtranslated sentences from SentNoB dataset

Dataset	Text	Assigned	Correct
VITD	আপু আপনি একটু হিজাপ পরেন। ধন্যবাদ	Direct Violence	Non Violence
	সঠিক কাজ হয়েছে। শিক্ষা প্রতিষ্ঠান ধর্ম পালনের জায়গা না। এখানে শিক্ষা ও ভবিষ্যৎ গড়ার লক্ষ্যে আসে। বাংলাদেশের কোন হাইস্কুলে বোরকা পড়ার বিধান আছে?	Direct Violence	Passive Violence
	অভিযোগ আবার কেমন শব্দ নষ্টা? হুতাশে এই পেশায় চলে এসেছিস নাকি? গুলি করেছে বলবি!!	Non Violence	Direct Violence
	নাটক টা অনেক ভাল ছিল 😒	Passive Violence	Non Violence
	মুসলিমদের জন্য আলাদা শিক্ষা প্রতিষ্ঠান করা হোক।।	Direct Violence	Non Violence
BLP Task 2	খুব ভালো সিদ্ধান্ত	Neutral	Positive
	পরিবেশবান্ধব পদ্ধতিতে জুমচাষের পরামর্শ বিশেষজ্ঞদের	Negative	Neutral
	আইপিএলের ইতিহাসে নতুন মাইলফলক স্থাপন করলেন কোহলি	Negative	Neutral
	দুঃখ জনক ঘটনা	Positive	Neutral
	করোনায় আরও ৩০ জনের মৃত্যু , মোট প্রাণহানি ৩১৮৪	Positive	Neutral
SentNoB	তুমি রেপারই হও , ডাক্তার হওয়ার দরকার নাই তোমার	Neutral	Negative
	এবার হিজরাদের নিয়ে রিপোর্ট করেন ওদের যন্ত্রণায় মানুষ অতিষ্ঠ হয়ে আছে	Positive	Negative

Figure 3: Examples of few incorrect labels we identified in the datasets. 'Assigned' shows the original label and 'Correct' shows our assessment of the appropriate label, illustrating label noise across all classes.

semantic-consistency thresholds, uncertainty-aware sampling) are orthogonal enhancements and left to future work.

3.3 Datasets

For our tasks, we utilized four datasets: BLP Shared Task 1: Violence Inciting Text Detection (VITD)(Saha et al., 2023), BLP Workshop Shared Task 2(Saha et al., 2023), SentNoB (Islam et al., 2021), and BD-SHS(Romim et al., 2022).

These datasets were selected to provide diverse evaluation scenarios for augmentation techniques. BD-SHS contains over 50,200 binary-labeled comments (hate/non-hate) from social media, serving as a control for binary classification against multi-class tasks. SentNoB comprises 15,000 manually annotated samples from social media across 13 domains with three-way sentiment labels (positive, negative, neutral), representing noisy informal Bangla text. VITD focuses on violence detection with three classes (direct violence, passive violence, non-violence) from YouTube comments about violent incidents in Bengal. BLP Task 2 involves three-class sentiment analysis of social media posts. This diversity in task types (hate speech detection, violence detection, sentiment analysis), class configurations (binary vs. ternary), and data characteristics (formal vs. noisy text) enables comprehensive assessment of how dataset properties affect augmentation efficacy.

These datasets serve as reference points for evaluating the effectiveness of the augmentation techniques. Here, BD-SHS serves more as a control for binary classification against the more complex multi-class classification. Apart from that, as stated earlier, we only augmented the minority class samples. For SentNoB and BLP Task 2, the minority class was the neutral class and for VITD it was the direct violence class (15%). Our augmentation results can be seen in Figure 1. Through our augmentation we mostly doubled the samples of minority class in order to improve the class imbalance of each of the datasets. We show the quality of backtranslation and paraphrased sentences in Figure 2. These sentences were taken from SentNoB dataset.

Additionally, as mentioned earlier, we found considerable label noise in the datasets from our qualitative analysis and we showcase some of these in Figure 3. We can see that samples across all classes are mislabeled and this is likely to impact the performance of models as this increases both false positive and false negative rates. We touch upon the impact of label noise on performance in Section 4.3.

Code and recipes. All augmentation scripts, preprocessing, and training configurations used in this study are available at this repository.

3.4 Experiment Setup

As described in the previous section, for the experiment setup, two additional versions of each dataset were produced using (Akil et al., 2022) for paraphrasing and (Hasan et al., 2020) for backtransla-

Dataset	Version	mBERT				XLM-Indic				BanglaBert			
		F1-Macro	F1-Micro	Recall	Prec.	F1-Macro	F1-Micro	Recall	Prec.	F1-Macro	F1-Micro	Recall	Prec.
BD-SHS	Baseline	91.9648	**91.9666**	91.9886	92.0494	91.8253	91.8274	91.8441	91.9061	91.9680	91.9666	91.8753	91.8713
	Paraphrased	91.4841	91.4894	91.4733	91.5325	92.3445	**92.3444**	92.4433	92.4692	92.3443	**92.3444**	92.4433	92.3444
	Backtranslated	90.9909	90.9922	90.9922	91.0293	90.5150	90.5150	90.6624	90.5150	91.4295	91.4297	91.6191	91.5874
SentNoB	Baseline	**69.5592**	71.6267	69.5000	69.7422	**68.3493**	70.4288	69.7352	69.7352	69.4012	73.7705	70.3164	69.3870
	Paraphrased	67.3360	69.1677	67.4392	67.6137	67.2832	69.4830	67.3433	67.4474	**72.2954**	75.8512	73.0009	72.1994
	Backtranslated	66.6228	68.1589	67.0720	67.1482	68.2108	70.1135	68.3326	68.5054	72.0372	75.5359	72.6263	71.8748
VITD	Baseline	**65.8050**	70.8333	64.9746	69.2229	66.4206	71.8254	65.1105	68.3673	72.3954	76.8353	71.1721	76.4834
	Paraphrased	65.0452	70.1389	64.1128	69.8489	65.0522	70.6845	65.1723	69.3623	**74.6663**	78.6210	73.7993	77.7827
	Backtranslated	64.1207	69.5933	63.8942	68.2788	**68.1876**	70.2396	68.1758	68.4270	74.3087	78.1746	72.6043	78.9988
BLP Task 2	Baseline	58.6598	**61.0109**	58.6969	60.2731	61.9487	**65.0962**	61.5665	62.8359	66.4386	71.3434	66.2780	66.6474
	Paraphrased	56.4873	59.3112	56.8890	57.4954	59.8386	63.7841	59.53085	60.3757	66.6997	**72.1403**	70.9197	72.1403
	Backtranslated	57.3038	59.9374	57.4055	58.5438	60.4425	63.8288	60.16027	61.2324	64.5253	70.2550	64.9660	64.5774

orange indicates the best performing augmentation method for each model and blue indicates the best performing model within each dataset.

Table 1: Evaluation of models on different datasets and their augmentations

tion. We augmented only the minor class in the training split. It should be noted that the dataset BD-SHS (Romim et al., 2022) is not imbalanced. Nevertheless, we still augmented it to increase its size to twice its original volume.

For our experiments, we used three different models. Here, our main goal was to compare the results between monolingual, language family specific multilingual and general multilingual language models. Hence, we used BanglaBERT (Bhattacharjee et al., 2021), as our monolingual language model, XLM-Indic (Moosa et al., 2023), as our language family specific multilingual language model and finally mBERT (Devlin et al., 2018), as our multilingual model with large number of languages.

All the models were trained for 3 epochs. All the baseline models except XLM-Indic had a learning rate of 2e-5.The learning rate for XLM-Indic was set at 3e-5 for the BLP Task 2 dataset and 4.5e-5 for rest of the baseline datasets. Furthermore, the learning rates required for the augmented datasets were higher and ranged from 3e-5 to 8e-5. Further details on hyperparameters are provided in Appendix A.

3.5 Hyperparameter Rationale

The hyperparameter selection was guided by model and dataset characteristics. All models were trained for 3 epochs, which proved sufficient for convergence without overfitting. Baseline models generally performed well with a learning rate of 2e-5, standard for BERT fine-tuning. However, XLM-Indic required higher learning rates (3e-5 to 4.5e-5) for baseline datasets, likely due to its multilingual pre-training requiring more aggressive updates to adapt to Bangla-specific sentiment patterns.

Augmented datasets consistently required higher learning rates (3e-5 to 8e-5) compared to baselines. This increase was necessary because augmentation introduced greater variance through paraphrasing and backtranslation, requiring models to adapt to a wider distribution of linguistic expressions. The doubled dataset size also necessitated more aggressive gradient steps for convergence within the same epochs. BanglaBERT, being monolingual, showed more stability with moderate learning rate increases, while multilingual models (mBERT and XLM-Indic) needed more varied adjustments to effectively utilize the augmented samples.

4 Results and Discussion

In this study, we explored how generative data augmentation can potentially enhance performance in semantic classification task for Bangla. In the subsequent section, we present the results of our findings and also show how the improvements from augmentations can be affected in presence of noisy labels.

4.1 Data augmentation improves performance

From Table 1, we can see that data augmentation indeed improves the performance of models. However, this improvement is not entirely universal. Here we will discuss from both model and augmentation technique point of view. From augmentation technique side of things, we can observe that as a technique, paraphrasing outperforms backtranslation. For instance, BanglaBert sees an improvement of approximately 2% F1-Micro for BD-SHS, 3% F1-Macro for SentNoB, 3% F1-Micro for VITD and finally 1% F1-Micro on BLP Task 2. Although backtranslation did not show improve-

Dataset	Classes	F1-Macro	Accuracy	Recall	Prec.
SentNoB	No Neutral	90.57	90.61	90.58	90.55
	No Positive	79.19	80.79	80.39	78.52
	No Negative	74.12	77.54	76.26	73.14
BLP Task 2	No Neutral	83.88	84.57	83.59	84.28
	No Positive	70.24	76.97	71.06	69.62
	No Negative	75.42	78.24	78.36	74.40
VITD	No Passive Violence	88.77	93.75	86.78	91.23
	No Direct Violence	79.42	81.32	82.35	78.39
	No Non Violence	78.25	82.61	76.14	84.03

Table 2: Impact of label noise on performance

ment in all the datasets for Banglabert, we can still see that it improved the baseline by 2.6% for SentNoB dataset. Apart from that, backtranslation also improved the baseline results of XLM-Indic on VITD by 2% and paraphrase improves BD-SHS baseline by 0.5%. On average paraphrasing gave the best improvements compared to backtranslation. Given the texts are from social media and are noisy by nature, it might be that backtranslation further added some noise due to translation process whereas paraphrasing might have had better noise to variance ratio due to its permutation nature. Another interesting thing to observe is the BD-SHS dataset. Most models performed quite well on it and this is likely due to its classes polar and binary nature. Hence, we did not see as much improvement from augmentation and the results of XLM-Indic and BanglaBert also seems to match for the paraphrased BD-SHS.

4.2 Impact of Pretraining

From a pretraining perspective, we see that the monolingual model performed much better with augmentations compared to the multilingual models. In most cases the multilingual models performed worse than their baseline results with the augmented datasets. This is likely due to the noise to variance ratio we discussed earlier. It seems that, monolingual models are better suited to use the added variance of the augmentation methods compared to multilingual models. This finding is not in line with (Ghosh and Senapati, 2022) where they report multilingual model perform as well as monolingual models on a similar task in Bangla.

4.3 Label Noise Impedes Performance Gains

In Section 4.1, we saw how data augmentation improved performance over baseline. However, this performance could be improved much more with better data quality. For instance, in Figure 3, we showed how these datasets have noisy labels and in some cases really poor inter annotator agreement (Islam et al., 2021) and how these noisy labels may impact performance. Here, we show empirically that, noisy label induced ambiguous decision boundary indeed degrades performance. We show this by performing binary classification on the baseline datasets using BanglaBert. Our main hypothesis here was that, the performance difference between our binary classifiers would not be drastically high if they were represented equally without label noise. However, in presence of label noise, the class boundary would be ambiguous and that would degrade the model. We can exactly see this in Table 2. Here, we can see that on SentNoB, removing the neutral class results in the best performing model. Whereas, in presence of neutral class, we see a 10% reduction of accuracy score. Specially, we can observe that the classifier for positive and neutral classes performed the worst out of all three permutations. It is expected that polarized classes like positive and negative would be easier to learn and neutral classes being somewhat in the middle might be harder to learn. However, here we see considerable degradation. Hence, we believe that the positive class has more overlap with the neutral class and as stated earlier, this results in poorer decision boundaries for classifiers. The baseline result for SentNoB was 69.40 F1-Macro and the best result was paraphrased SentNoB with a score of 72.30 F1-macro. Comparing these to the results in Table 2 can give us some idea how the results degrade due to noisy labels. Furthermore, we can see similar trends for both VITD and BLP Task 2. Both of them show on average almost 10% reduction in accuracy score. As discussed earlier, BLP Task 2 is a mix of SentNoB and MUBASE dataset. Hence, similar to results on 1, it gives us a glimpse of the error propagation

of MUBASE. We can see that compared to 90.57 accuracy score of SentNoB's no neutral class subset, the no neutral class subset of BLP Task2 has an accuracy score of 84.57. Again we see that the main issue here is the neutral class. This leads us to believe that neutral class annotation are the main cause of label noise and it requires better attention.

4.4 Issues with Neutral Sentiment

The distinction between neutral and other sentiments is where the model struggles the most, for both positive and negative. Unlike polarized classes, which may have specific lexical indicators, the neutral class lacks such clear markers. We hypothesize that the neutral classes have a much higher distribution than the polarized classes. To accurately represent the distribution of these neutral classes, we recommend a much higher representation of neutral class than the polarized classes.

We would also advocate for the introduction of an "indeterminate" class to address another issue. While a neutral sentiment refers to an unpolarized yet clearly determined sentiment, an indeterminate label captures instances where sentiment is genuinely unclear or ambiguous. Whether this deserves its own category requires further analysis and validation.

By adopting this labeling scheme, we can ensure that the model does not mistakenly categorize uncertain sentiments into the neutral category, thereby preserving the integrity of both classes.

5 Limitations

We focus on encoder-only models to control confounds across monolingual vs. multilingual settings under matched budgets, leaving decoder-only and sequence-to-sequence architectures to future work. While we emphasize sentence-level augmentation for semantic fidelity, a broader comparison with additional augmentation families (e.g., mixup/noising) is also deferred.

6 Conclusion

In this study, we demonstrate that data augmentation techniques, notably paraphrasing and backtranslation, enhance the performance of Bangla sentiment classifiers. Our results reveal that paraphrasing significantly benefits monolingual models, more so than backtranslation does for multilingual models. We also detected noisy labels across all four datasets. Our analysis provides empirical evidence that label noise hampers classifier performance, with the neutral class emerging as the primary source of this noise. Given these findings, we argue for robust protocols for annotating neutral classes. We propose weighting inter-annotator agreement by class, suggesting the neutral class be assigned the highest weight. Consequently, the neutral class should attain higher inter-annotator agreement scores compared to the positive and negative classes.

References

S. M. Abu Taher, Kazi Afsana Akhter, and K. M. Azharul Hasan. 2018. N-gram based sentiment mining for Bangla text using support vector machine. In *Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. *arXiv preprint arXiv:2210.05109*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Anirban Bhowmick and Abhik Jana. 2021. Sentiment analysis for Bengali using transformer based models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 481–486, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, and M. Rubaiyat Hossain Mondal. 2022. Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, 13:100123.

Partha Chakraborty, Farah Nawar, and Humayra Afrin Chowdhury. 2022. A ternary sentiment classification of Bangla text data using support vector machine and random forest classifier. In Jyotsna Kumar Mandal, Pao-Ann Hsiung, and Rudra Sankar

Dhar, editors, *Topical Drifts in Intelligent Computing*, pages 69–77. Springer Nature Singapore, Singapore.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Koyel Ghosh and Dr. Apurbalal Senapati. 2022. Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 853–865, Manila, Philippines. De La Salle University.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. *arXiv preprint arXiv:2009.09359*.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2020. Sentilstm: a deep learning approach for sentiment analysis of restaurant reviews. In *International Conference on Hybrid Intelligent Systems*, pages 193–203. Springer.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Khondoker Ittehadul Islam, Md Saiful Islam, and Md Ruhul Amin. 2020. Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually augmented data. In *International Conference on Learning Representations*.

Mst Eshita Khatun and Tapasy Rabeya. 2022. A machine learning approach for sentiment analysis of book reviews in bangla language. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1178–1182. IEEE.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Forhad An Naim. 2021. Bangla aspect-based sentiment analysis based on corresponding term extraction. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 65–69.

Md Rafi-Ur-Rashid, Mahim Mahbub, and Muhammad Abdullah Adnan. 2022. Breaking the curse of class imbalance: Bangla text classification. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–21.

Shadman Rohan, Mojammel Hossain, Mohammad Mamun Or Rashid, and Nabeel Mohammed. 2023. Bencoref: A multi-domain dataset of nominal phrases and pronominal reference annotations. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW)*.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5153–5162, Marseille, France. European Language Resources Association.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Blp-2023 task 1: Violence inciting text detection (vitd). In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Md Ferdous Wahid, Md Jahid Hasan, and Md Shahin Alom. 2019. Cricket sentiment analysis from bangla text using recurrent neural network with long short term memory model. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33. Curran Associates, Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Finetuning Hyperparameters

Dataset	Batch Size	Learning Rate	Weight Decay	Dropout	Epochs	Warmup Ratio	Label Smoothing
BD-SHS Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
BD-SHS Paraphrased	32	5e-5	1e-6	0.15	3	0.10	0.15
BD-SHS Backtranslated	32	6e-5	1e-3	0.15	3	0.10	0.15
SentNoB Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
SentNoB Paraphrased	32	5.5e-5	1e-3	0.15	3	0.10	0.15
SentNoB Backtranslated	32	2e-5	1e-3	0.15	3	0.10	0.15
VITD Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
VITD Paraphrased	32	3e-5	1e-3	0.15	3	0.10	0.15
VITD Backtranslated	32	6e-5	1e-3	0.15	3	0.10	0.15
BLP Task 2 Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
BLP Task 2 Paraphrased	32	3e-5	1e-6	0.15	3	0.10	0.15
BLP Task 2 Backtranslated	32	6e-5	1e-3	0.15	3	0.10	0.15

Table 3: Hyperparameters for mBERT model

Dataset	Batch Size	Learning Rate	Weight Decay	Dropout	Epochs	Warmup Ratio	Label Smoothing
BD-SHS Baseline	32	4.5e-5	1e-6	0.15	3	0.10	0.15
BD-SHS Paraphrased	32	5e-5	1e-3	0.15	3	0.10	0.15
BD-SHS Backtranslated	32	6e-5	1e-6	0.15	3	0.10	0.15
SentNoB Baseline	32	4.5e-5	1e-6	0.15	3	0.10	0.15
SentNoB Paraphrased	32	6e-5	1e-6	0.15	3	0.10	0.15
SentNoB Backtranslated	32	3.5e-5	1e-6	0.15	3	0.10	0.15
VITD Baseline	32	4.5e-5	1e-6	0.15	3	0.10	0.15
VITD Paraphrased	32	5.5e-5	1e-3	0.15	3	0.10	0.15
VITD Backtranslated	32	4e-5	1e-6	0.15	3	0.10	0.15
BLP Task 2 Baseline	32	3e-5	1e-6	0.15	3	0.10	0.15
BLP Task 2 Paraphrased	32	4e-5	1e-6	0.15	3	0.10	0.15
BLP Task 2 Backtranslated	32	3e-5	1e-6	0.15	3	0.10	0.15

Table 4: Hyperparameters for XLM-Indic model

Dataset	Batch Size	Learning Rate	Weight Decay	Dropout	Epochs	Warmup Ratio	Label Smoothing
BD-SHS Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
BD-SHS Paraphrased	32	5e-5	1e-3	0.15	3	0.10	0.15
BD-SHS Backtranslated	32	5e-5	1e-3	0.15	3	0.10	0.15
SentNoB Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
SentNoB Paraphrased	32	5e-5	1e-3	0.15	3	0.10	0.15
SentNoB Backtranslated	32	8e-5	1e-3	0.15	3	0.10	0.15
VITD Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
VITD Paraphrased	32	5e-5	1e-3	0.15	3	0.10	0.15
VITD Backtranslated	32	4e-5	1e-3	0.15	3	0.10	0.15
BLP Task 2 Baseline	32	2e-5	1e-3	0.15	3	0.10	0.15
BLP Task 2 Paraphrased	32	2e-5	1e-3	0.15	3	0.10	0.15
BLP Task 2 Backtranslated	32	5e-5	1e-3	0.15	3	0.10	0.15

Table 5: Hyperparameters for BanglaBERT model